Claims

1. In a communication system network having a plurality of servers, each of said plurality of servers having a load level based on serving a number of clients in said communication system network, a method comprising the steps of:

grouping said plurality of servers into a first and second server groups, wherein said first server group has a load level less than load level of said second server group;

calculating a time period T;

assigning load to a server selected from servers in said first server group from an initial time until expiration of said time period T;

assigning load to a server selected from servers in said first and second server groups after expiration of said time period T.

2. In a communication system network having a plurality of servers, each of said plurality of servers having a load level based on serving a number of clients in said communication system network, a method comprising the steps of:

grouping said plurality of servers into a plurality of server groups G0 through G2, wherein server groups G0 through G2 respectively have load levels progressively from a least amount of load level to a most amount of load level;

calculating time periods T1 and T2, wherein said time period T2 is longer than said time period T1;

assigning load to a server selected from servers in said server group G0 from an initial time until expiration of said time period T1;

assigning load to a server selected from servers in said server groups G0 and G1 after expiration of said time period T1;

assigning load to a server selected from servers in said server groups G0, G1 and G2 after expiration of said time period T2.

3. In a communication system network having a plurality of servers, each of said plurality of servers having a load level based on serving a number of clients in said communication system network, a method comprising the steps of:

grouping said plurality of servers into a plurality of server groups G0 through Gk, wherein said server groups G0 through Gk respectively have load levels progressively from a least amount of load level to a most amount of load level;

calculating a plurality of time periods T1 through Tk;

assigning load to a server selected from servers in said server group G0 from an initial time until expiration of said time period T1;

assigning load, after expiration of each of said time periods T1 through Tk measured from said initial time, to a server selected from the servers in the server groups from G0 and at least one other server group selected from said server groups G1 through Gk.

4. In a communication system network having a plurality of servers, each of said plurality of servers having a load level based on serving a number of clients in said communication system network, a method comprising the steps of:

grouping said plurality of servers into a plurality of server groups G0 through Gk, wherein said server groups G0 through Gk respectively have load levels from progressively a least amount of load level to a most amount of load level;

calculating a plurality of time periods T1 through Tk corresponding to said server groups G1 through Gk;

assigning load to a server selected from servers in said server group G0 from an initial time until expiration of said time period T1;

assigning load, after expiration of each of said time periods T1 through Tk measured from said initial time, to a server selected from servers in a combination of servers including said server group G0 and at least one other server group, in said server groups G1 through Gk, corresponding to an expiring time period.

5. The method as recited in claim 4 wherein said plurality of time periods T1 through Tk each is based on a difference between load levels of at least two server groups in said plurality of server groups G0 through Gk.

6. The method as recited in claim 4 further comprising the step of:

receiving an update of load level of at least one of said plurality of servers in said plurality of server groups G0 through Gk;

repeating said grouping to produce a new plurality of server groups G0 through Gk based on said update of load level;

repeating said calculating said plurality of time periods to produce a new plurality of time periods T1 through Tk corresponding to said new plurality of server groups G0 through Gk;

resetting said initial time to a reset initial time, and assigning load to a server selected from servers in said new server group G0 from said reset initial time until expiration of said new time period T1;

assigning load, after expiration of each of said new time periods T1 through Tk measured from said reset initial time, to a server selected from servers in a combination of servers including said new server group G0 and at least one other server group, in said new server groups G1 through Gk, corresponding to an expiring time period.

7. The method as recited in claim 4 wherein said grouping of said plurality of server groups G0 through Gk is based on similarity of load levels among said plurality of servers.

8. The method as recited in claim 4 wherein at least one load assignment in said assigning load to a server in said server group G0 and said assigning load to a server selected from servers in said combination is performed according to a round robin selection method.

9. The method as recited in claim 4 wherein at least one load assignment in said assigning load to a server in said server group G0 and said assigning load to a server selected from servers in said combination is performed according to a random selection method.

10. The method as recited in claim 4 wherein each of said plurality of time periods T1 through Tk is based on load levels of at least two server groups selected from said plurality of server groups G0 through Gk, a request arrival rate and a server service rate.

11. The method as recited in claim 10 wherein said request arrival rate is substituted for an average request arrival rate of said plurality of servers.

12. The method as recited in claim 10 wherein said request arrival rate is substituted for an average request arrival rate of a combination of servers of said plurality of servers.

13. The method as recited in claim 10 wherein said server service rate is substituted for an average service rate of said plurality of servers.

14. The method as recited in claim 10 wherein said server service rate is substituted for an average service rate of a combination of servers of said plurality of servers.